Статистика ФИВТ ПМИ

Прикладной поток

Лекция 11

5. Проверка статистических гипотез

5.4. p-value

Гипотезы (напоминание)

$X = (X_1, ..., X_n)$ — выборка из неизвестного распределения $\mathsf{P} \in \mathscr{P}$.

$\mathsf{H}_0 : \mathsf{P} \in \mathscr{P}_0$ — основная гипотеза;

$\mathsf{H}_1 : \mathsf{P} \in \mathscr{P}_1$ — альтернативная гипотеза.

$S \subset \mathscr{X}$ — критерий уровня значимости α для проверки H_0 *vs.* H_1,

если $\mathsf{P}(X \in S) \leqslant \alpha, \quad \forall\, \mathsf{P} \in \mathscr{P}_0$.

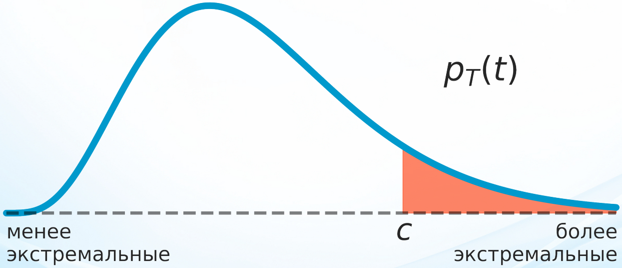
Варианты ответа:

1. $X \in S \quad \Longrightarrow \quad H_0$ отвергается $\Longrightarrow$ результат стат. значим;
2. $X \notin S \quad \Longrightarrow \quad H_0$ **не отвергается** $\Longrightarrow$ результат не стат. значим

Гипотезы (напоминание)

Часто критерий имеет вид $S = \{T(x) \geqslant c\}$,

где $T(X)$ — статистика критерия.



H_0 отвергается $\Longleftrightarrow$ $T(X) \geqslant c_\alpha$

Значение t_1 более экстремально, чем t_2, если $t_1 > t_2$.

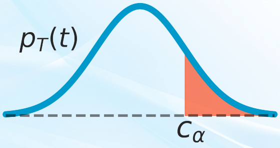
Критерии (напоминание)

Часто критерий имеет вид $S = \{T(x) \geqslant c_\alpha\}$,
где $T(X)$ — статистика критерия.

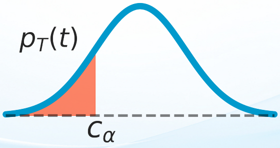
α выбирается **ДО** эксперимента,

c_α вычисляется из условия $\mathrm{P}_0(T(X) > c_\alpha) \leqslant \alpha$.

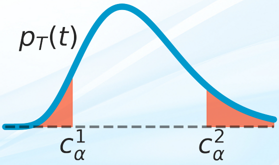
$S = \{T(x) > c_\alpha\}$



$S = \{T(x) < c_\alpha\}$



$S = \{|T(x)| > c_\alpha\}$



Замечание. Выбирать α после эксперимента неправильно.
Так можно подогнать результат под желаемый.
"*Статистика может доказать что угодно, даже истину.*"

Пример (влияние нового препарата на выздоровление)

Испытуемые делятся случайно на две группы:

1. *Исследуемая группа* — принимает новый препарат;

   $X = (X_1, \ldots, X_n), X_i \sim Bern(p_1)$ — результаты лечения.

2. *Контрольная группа* — принимает плацебо;

   $Y = (Y_1, \ldots, Y_m), Y_i \sim Bern(p_2)$ — результаты лечения.

$\mathrm{H}_0: p_1 = p_2$ — отсутствие эффекта

$\mathrm{H}_1: p_1 > p_2$ — эффект присутствует

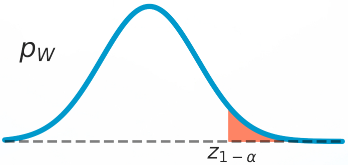
Пример (влияние нового препарата на выздоровление)

$\widehat{p}_1 = \overline{X} \overset{d}{\approx} \mathcal{N}\left(p_1, \frac{p_1(1-p_1)}{n}\right)$ и $\widehat{p}_2 = \overline{Y} \overset{d}{\approx} \mathcal{N}\left(p_2, \frac{p_2(1-p_2)}{m}\right)$ — ОМП

При справедливости H_0 получаем

$$W(X, Y) = \frac{\widehat{p}_1 - \widehat{p}_2}{\widehat{\sigma}} \overset{d}{\approx} \mathcal{N}(0, 1),$$

где $\widehat{\sigma} = \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{m}}$.



// Сходимость $W(X, Y) \xrightarrow{d} \mathcal{N}(0, 1)$ при $n, m \to +\infty$ можно доказать строго.

Критерий Вальда $S = \{W(x, y) > z_{1-\alpha}\}$.

$\alpha = 0.05 \quad \Longrightarrow \quad z_{1-\alpha} \approx 1.64, \quad S = \{W(x, y) > 1.64\}$.

Дов. интервал для $p_1 - p_2$ равен $C = (\widehat{p}_1 - \widehat{p}_2 - z_{1-\alpha}\widehat{\sigma},\ 1)$.

H_0 отвергается $\Longleftrightarrow 0 \notin C$

Пример (влияние нового препарата на выздоровление)

1. 1 группа: $n = 30$ человек, 27 выздоровело $\Longrightarrow$ $\widehat{p}_1 = 0.9$

   2 группа: $m = 30$ человек, 21 выздоровело $\Longrightarrow$ $\widehat{p}_2 = 0.7$

   $W(x, y) \approx 2$ $\Longrightarrow$ H_0 отвергается, результат стат. значим

   дов. интервал $(0.036, 1)$ $\leftarrow$ **слабая уверенность в результате**

2. 1 группа: $n = 30$ человек, 27 выздоровело $\Longrightarrow$ $\widehat{p}_1 = 0.9$

   2 группа: $m = 30$ человек, 15 выздоровело $\Longrightarrow$ $\widehat{p}_2 = 0.5$

   $W(x, y) \approx 3.76$ $\Longrightarrow$ H_0 отвергается, результат стат. значим

   дов. интервал $(0.225, 1)$ $\leftarrow$ **хорошая уверенность в результате**

3. 1 группа: $n = 30$ человек, 27 выздоровело $\Longrightarrow$ $\widehat{p}_1 = 0.9$

   2 группа: $m = 10$ человек, 7 выздоровело $\Longrightarrow$ $\widehat{p}_2 = 0.7$

   $W(x, y) \approx 1.54$ $\Longrightarrow$ H_0 не отвергается, результат стат. незнач.

   дов. интервал $(-0.017, 1)$ $\leftarrow$ **нет результата**

Материал на доске

p-value (достигаемый уровень значимости)

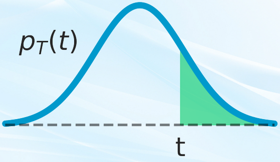
$H_0 : \mathrm{P} \in \mathscr{P}_0$ $\qquad$ $x_1, \ldots, x_n$ — реализация выборки

$T(x)$ — статистика критерия $\qquad$ $t = T(x_1, \ldots, x_n)$ — реализация стат.

p-value — вероятность получить при справедливости H_0 такое значение статистики $t = T(x)$ или еще более экстремальное.

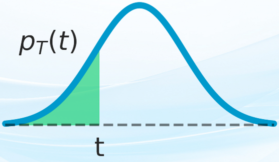
$S = \{T(x) > c_\alpha\}$

$p(x) = \mathrm{P}_0(T(X) \geqslant t)$,



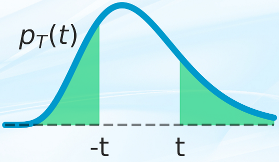
$S = \{T(x) < c_\alpha\}$

$p(x) = \mathrm{P}_0(T(X) \leqslant t)$,



$S = \{|T(x)| > c_\alpha\}$

$p(x) = \mathrm{P}_0(T(X) \geqslant |t|) + \mathrm{P}_0(T(X) \leqslant -|t|)$,

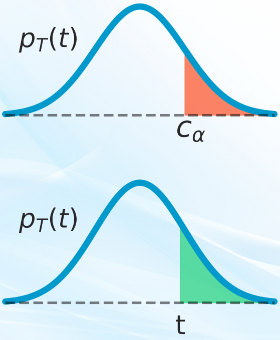


Замечание. Если распр. $T(X)$ при H_0 не одинаково, то нужно добавить $\sup\limits_{\mathrm{P} \in \mathscr{P}_0}$

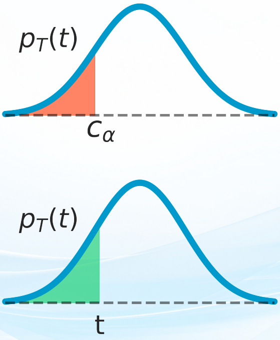
В чем же разница? Графики одинаковые!!!

Еще раз посмотрим на них:

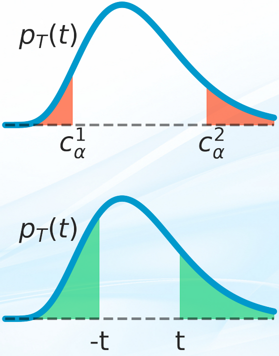
$S = \{T(x) > c_\alpha\}$



$S = \{T(x) < c_\alpha\}$



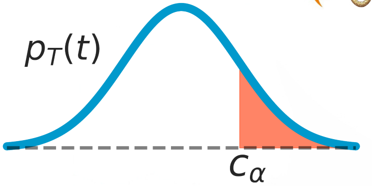
$S = \{|T(x)| > c_\alpha\}$

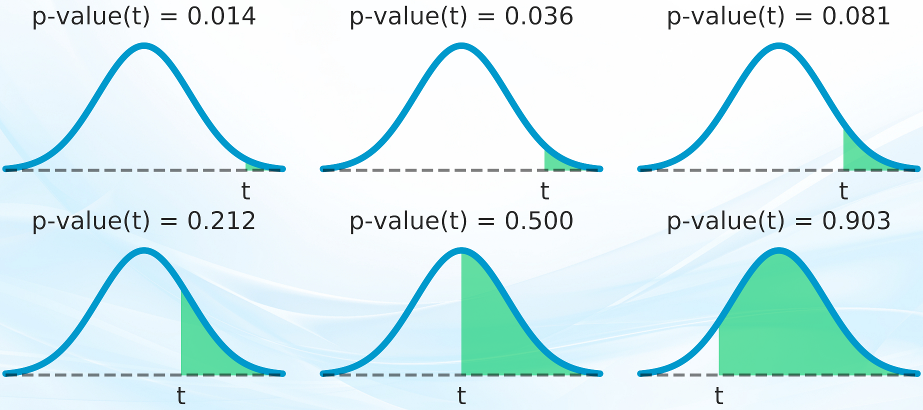


Рассмотрим случай $S = \{T(x) > c_\alpha\}$

Критическое множество (слева) фиксировано.

Ниже p-value для различных реализаций.





Вывод:

H_0 отвергается $\Longleftrightarrow$ p-value $\leqslant \alpha$

В этом случае p-value —
степень уверенности в отвержении H_0.

(чем p-value меньше, тем увереннее)

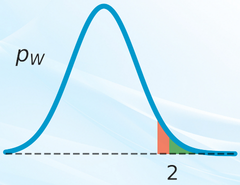
Пример (влияние нового препарата на выздоровление)

Критерий $S = \{W(x,y) > z_{1-\alpha}\}$, где $W(X,Y) \xrightarrow{d} \mathcal{N}(0,1)$.

p-value: $p(w) = \mathsf{P}(W(X,Y) > w) = \text{scipy.stats.norm.sf}(w)$.

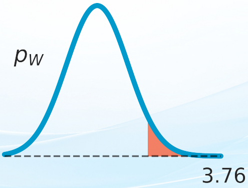
$w = W(x) = 2$

$p(w) = 0.0228$



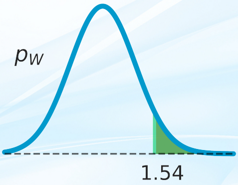
$w = W(x) = 3.76$

$p(w) = 0.00008$



$w = W(x) = 1.54$

$p(w) = 0.0618$



Свойство p-value

Утверждение.

Если при справедливости H_0

распр. статистики $T(X)$ одинаково и непрерывно,

то $p(T(X)) \sim U[0, 1]$ при H_0.

Замечание.

Часто на практике это верно, т.к. многие критерии так и строятся.

Следствие.

Значения 0.2 и 0.9 одинаковы с точки зрения справедливости H_0,

т.е. p-value не есть степень уверенности в справедливости H_0.

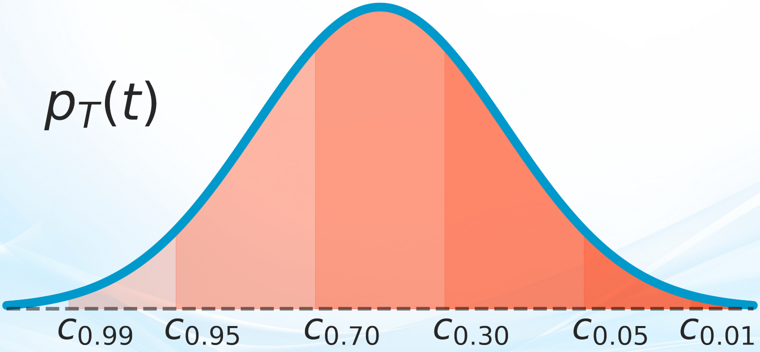
Что возможно, если p-value большой?

1. H_0 верна;
2. Критерий недостаточно мощный.

Общий случай p-value

$\{S_\alpha \mid \alpha \in [0,1]\}$ — семейство критериев для разных уровней значимости.

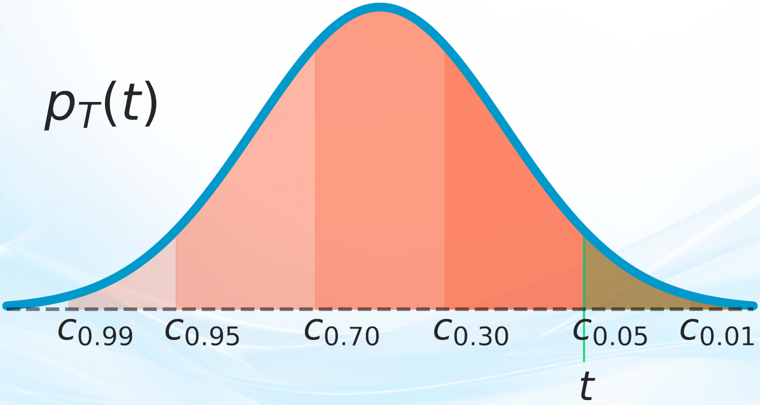
$S_\alpha = \{T(x) > c_\alpha\}$ — критерий



Общий случай p-value

$\{S_\alpha \mid \alpha \in [0,1]\}$ — семейство критериев для разных уровней значимости.

$S_\alpha = \{T(x) > c_\alpha\}$ — критерий

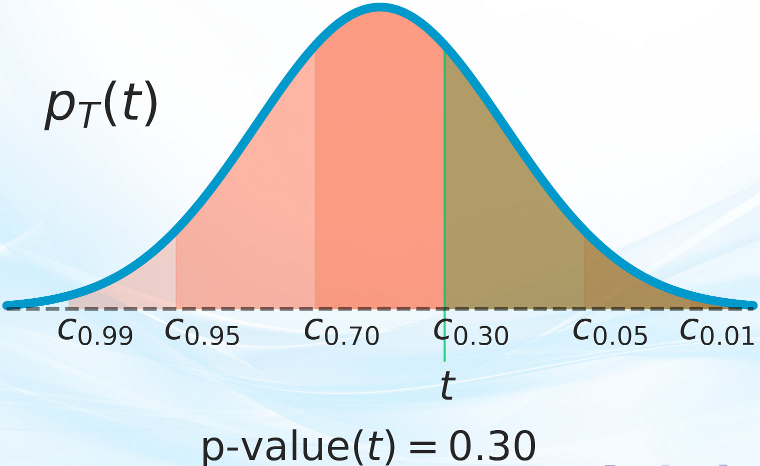


$$\text{p-value}(t) = 0.05$$

Общий случай p-value

$\{S_\alpha \mid \alpha \in [0,1]\}$ — семейство критериев для разных уровней значимости.

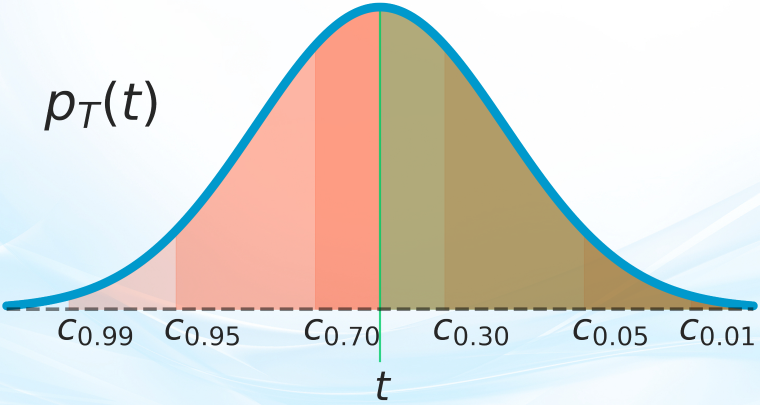
$S_\alpha = \{T(x) > c_\alpha\}$ — критерий



Общий случай p-value

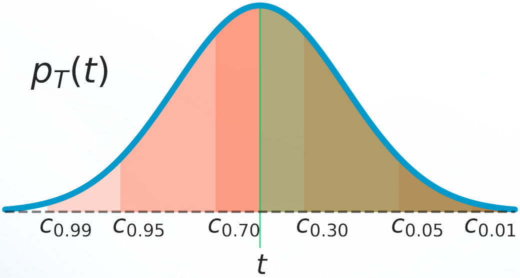
$\{S_\alpha \mid \alpha \in [0,1]\}$ — семейство критериев для разных уровней значимости.

$S_\alpha = \{T(x) > c_\alpha\}$ — критерий



p-value$(t) = 0.50$

Общий случай p-value



$$\text{p-value}(t) = 0.50$$

$t = c_{0.5} \implies$ при $\alpha \geqslant 0.5$ гипотеза H_0 отвергается.

при $\alpha < 0.5$ гипотеза H_0 не отвергается.

Ключевое наблюдение: Если отвергнуть H_0 можно только совершив большую ошибку, то скорее ее не стоит отклонять.

Общий случай p-value

Вывод:

p-value — наименьший уровень значимости, при котором H_0 можно отвергуть для данной реализации выборки x.

$$p(x) = \inf\{\alpha \in [0, 1] \mid x \in S_\alpha\}$$

Что не есть p-value

Величина p-value не является

- уровнем значимости, реальным уровнем значимости, вероятностью ошибки первого рода;

  не зависят от выборки

- вероятностью H_0, вероятностью H_0 при условии выборки;

  она либо верна, либо нет

- много чем еще.

Что не есть p-value (Пример)

На ЧМ по футболу в 2010 г. осьминог Пауль предсказывает результаты матчей с участием сборной Германии, выбирая кормушку с флагом страны-победителя.

$X_1, \ldots, X_n \sim Bern(p)$ — результаты предсказания (правильно/нет).

$\mathrm{H}_0 : p = 1/2$ *vs.* $\mathrm{H}_1 : p > 1/2$ (наугад *vs.* не наугад)

Критерий $S = \{T(x) > c_\alpha\}$, где $T(X) = \sum X_i \sim Bin(n, p)$.

p-value: $p(t) = \frac{1}{2^n} \sum\limits_{j=t}^{n} C_n^j$

13 матчей: Пауль верно угадывает исход матча в 11 случаях.

$p(11) = 2^{-13} \left(C_{13}^{11} + C_{13}^{12} + C_{13}^{13} \right) \approx 0.0112 < 0.05 \Longrightarrow \mathrm{H}_0$ отвергается;

0.0112 не является вероятность того, что Пауль выбирает кормушку наугад

Выводы:

$$H_0 \text{ отвергается} \iff \text{p-value} \leqslant \alpha$$

Если H_0 отвергается, то p-value — степень уверенности в отвержении H_0.

(чем p-value меньше, тем увереннее)

Если H_0 не отверг., то ничего сказать нельзя, p-value не есть степень уверенности в справ. H_0.

Если H_0 верна, то обычно $p(T(X)) \sim U[0,1]$.

5. Проверка статистических гипотез

5.5. Практическая значимость результата

Большие выборки

$X_1, ..., X_n \sim Bern(\theta)$ — результаты испытания Пауля.

$\mathrm{H}_0 : \theta = 1/2$ *vs.* $\mathrm{H}_1 : \theta > 1/2$

Критерий $S = \{T(x) \geqslant c_\alpha\}$, где $T(X) = \sum X_i \sim Bin(n, p)$.

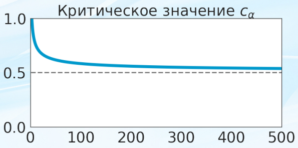
Как критическое значение c_α зависит от n?

Рассмотрим асимптотически эквивалентный критерий Вальда

$$W(X) = \sqrt{n}\frac{\overline{X} - 1/2}{\sqrt{1/4}} \xrightarrow{d_{1/2}} \mathcal{N}(0, 1).$$

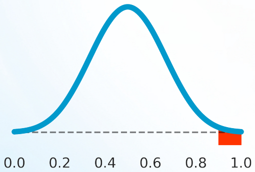
Тогда критерий

$$S_W = \{W(x) > z_{1-\alpha}\} = \left\{\overline{x} > \frac{1}{2} + \frac{z_{1-\alpha}}{2\sqrt{n}}\right\}.$$

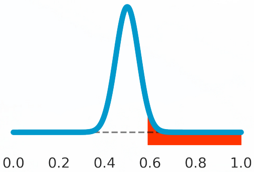


Большие выборки

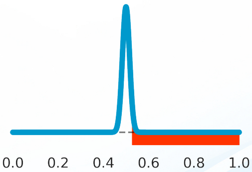
$n = 10, c_\alpha = 0.9$



$n = 100, c_\alpha = 0.59$



$n = 1000, c_\alpha = 0.527$



Вывод: при $n \to +\infty$ мощность критерия сходится к 1.

Теория: это замечательно!

Практика: не совсем...

Теория: даже при небольшом отличии истины от H_0 мы ее отклоним!

Практика: Хахаха, какой смысл в осьминоге, который угадывает с вероятностью 0.51? На сковородку его!

Вывод с точки зрения практики

Как правило, на практике:

1. **При малом размере выборки:**

   Почти ничего не отклоняется, т.к. мощность небольшая.

   Недообучение

2. **При большом размере выборки:**

   Отклоняется почти все, т.к. небольшие отличия от H_0 есть почти всегда.

   Переобучение

Практическая значимость

Размер эффекта — величина, оценивающая по данным, насколько основная гипотеза отличается от истины.

Пример 1.

В течении трех лет женщины выполняли физические упражнения в двух группах: группа 1 — не менее часа в день, группа 2 — не более 20 минут в день.

H_0: изменение веса в обоих группах одинаково

H_1: в 1 группе уменьшение веса больше, чем во 2-й

p-value $< 0.001 \implies$ результат статистически значим

Разница в весе **150 грамм** $\implies$ результат практически не значим

Практическая значимость

Размер эффекта — величина, оценивающая по данным, насколько основная гипотеза отличается от истины.

Пример 2.

В 2002 году проводились клинические испытания гормонального препарата Премарин, которые были досрочно прерваны.

На **0.08%** увеличивается риску развития рака груди;

На **0.08%** увеличивается риск инсульта;

На **0.07%** увеличивается риск инфаркта.

С учетом численности населения есть практическая значимость.

Практическая значимость

	Есть практ. значимость	Нет практ. значимости
Есть стат. значимость	H_0 отвергается: Эффект присутствует и доказан статистически	Скорее всего в полученном результате смысла нет
Нет стат. значимости	Эффект присутствует, но не доказан статистически; нужно продолжать эксперимент	H_0 не отвергается: результата нет

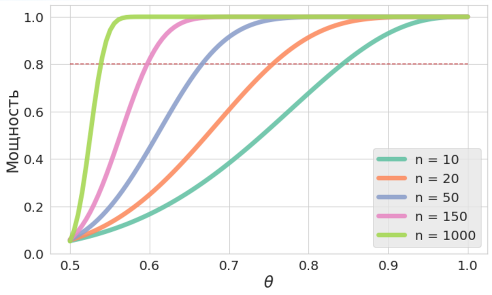
План эксперимента

Как определить размер выборки **до** эксперимента?

$X_1, \ldots, X_n \sim Bern(\theta)$

$\mathrm{H}_0 : \theta = 1/2$ *vs.* $\mathrm{H}_1 : \theta > 1/2$

Графики мощности для критерия $S = \{\sum X_i > c_\alpha\}$:



$\alpha = 0.05$ — ур. знач.

Желаемые значения:

$\beta = 0.8$ — мощность;

$\theta \geqslant 0.6$ — значимый эффект.

Выбираем n, для которого

кривая мощности проходит

через точку $(0.6,\ 0.8)$.

5. Проверка статистических гипотез

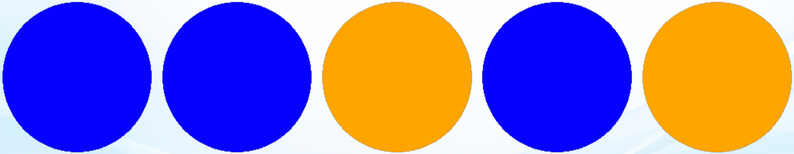
5.6. Множественная проверка гипотез

Поиск экстрасенсов

Этап 1: Угадайте цвета (**синий** и **оранжевый**) с учетом порядка.

Поиск экстрасенсов

Этап 1: ответы.

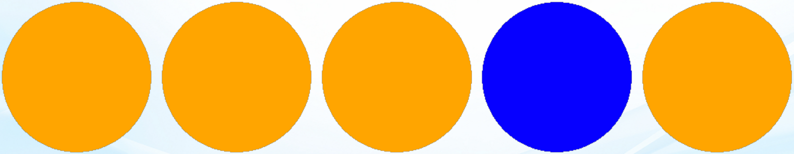


Поиск экстрасенсов

Этап 2: Угадайте цвета (**синий** и **оранжевый**) с учетом порядка.

Поиск экстрасенсов

Этап 2: ответы.



Поиск экстрасенсов

В 1950 г. проводились испытания

возможности экстрасенсорного восприятия.

Этап 1: поиск экстрасенсов — испытуемому нужно угадать цвет 10 карт.

$X_1, ..., X_{10} \sim Bern(\theta)$ — результаты (правильно / нет).

$\mathrm{H}_0 : \theta = 1/2$ *vs.* $\mathrm{H}_1 : \theta > 1/2$ (наугад *vs.* не наугад)

Критерий $S = \{T(x) \geqslant c_\alpha\}$, где $T(X) = \sum X_i \sim Bin(n, p)$.

c	7	8	9	10
$\mathrm{P}_{1/2}(T(X) \geqslant c)$	0.172	0.055	0.010	0.001

Берем $c_\alpha = 9$, т.е. H_0 отклоняется если $\sum X_i \geqslant 9$.

Поиск экстрасенсов

Вывод: если человек верно отгадывает хотя бы 9 карт из 10, то он становится предполагаемым экстрасенсом.

В эксперименте приняли участие 1000 человек, при этом

- 9 карт верно отгадали 9 человек;
- 10 карт верно отгадали 2 человека.

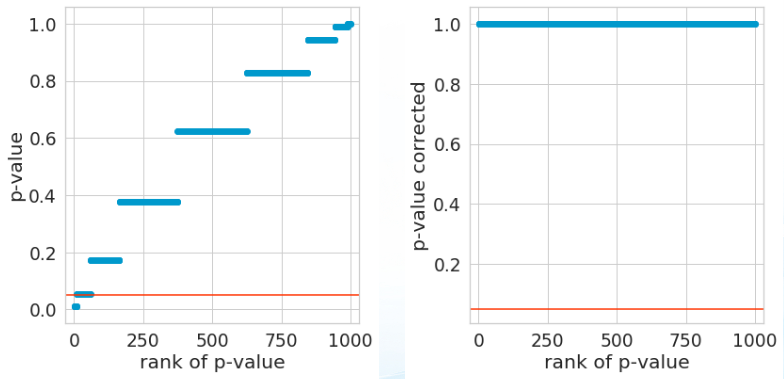
В дальнейшем ни один из них не подтвердил свои способности...

$$\mathrm{P}_{1/2}\,(\text{хотя бы один из 1000 угадает 9 или 10 карт верно}) =$$

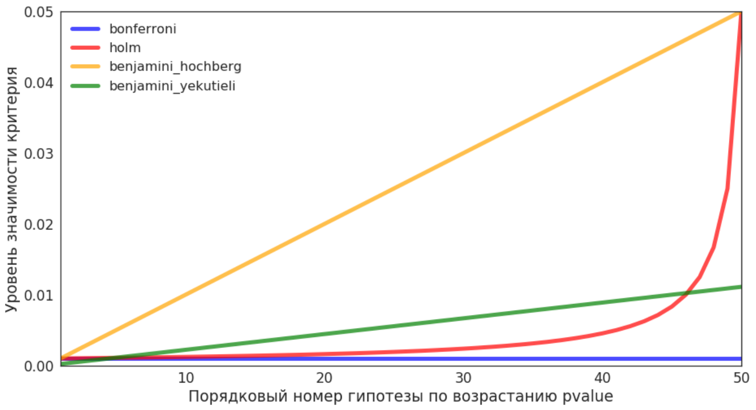
$$= 1 - \left(1 - C_{10}^{9}/2^{10} - C_{10}^{10}/2^{10}\right)^{1000} = 1 - \left(1 - 11/2^{10}\right)^{1000} \approx 0.99997$$

Материал на доске

Поиск экстрасенсов



Сравнение методов МПГ



Реализация МПГ

`statsmodels.stats.multitest.multipletests`

`(pvals, alpha=0.05, method='hs',`

`is_sorted=False, returnsorted=False)`

- `pvals` — значения p-value по всем критериям
- `alpha` — желаемый уровень значимости
- `method`:
  - `bonferroni`
  - `sidak`
  - `fdr_bh`
  - `holm`
  - `holm-sidak`
  - `fdr_by`

Возвращает:

- `reject` — для отвергаемых гипотез `True`
- `pvals_corrected` — скорректированые p-value

Простой пример

Знакомая задача:

$X_1, ..., X_n \sim \mathcal{N}(\theta, 1)$

$\mathrm{H}_0: \theta \geqslant 0$ *vs* $\mathrm{H}_1: \theta < 0$

РНМК: $S = \{x \in \mathbb{R} \mid \overline{x} \leqslant c_\alpha\}$

Пусть теперь две одинаковые задачи с независимыми выборками:

$X_1, ..., X_n \sim \mathcal{N}(\theta_1, 1)$

$Y_1, ..., Y_n \sim \mathcal{N}(\theta_2, 1)$

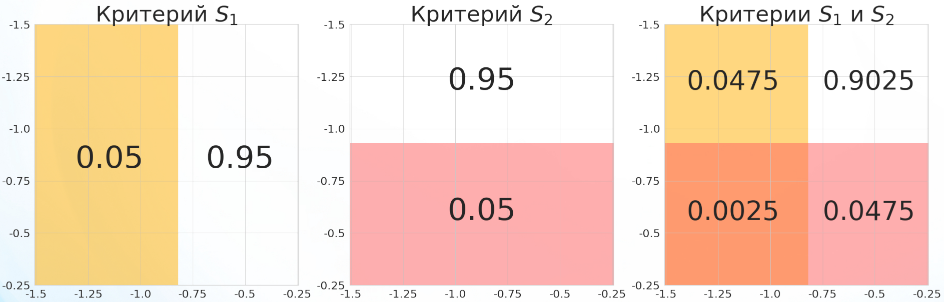
$\mathrm{H}_1: \theta_1 \geqslant 0$ *vs* $\mathrm{H}_1': \theta_1 < 0$

$\mathrm{H}_2: \theta_2 \geqslant 0$ *vs* $\mathrm{H}_2': \theta_2 < 0$

Критерии: $S_1 = \left\{(x, y) \in \mathbb{R}^2 \mid \overline{x} \leqslant c_\alpha\right\}$

$S_2 = \left\{(x, y) \in \mathbb{R}^2 \mid \overline{y} \leqslant c_\alpha\right\}$

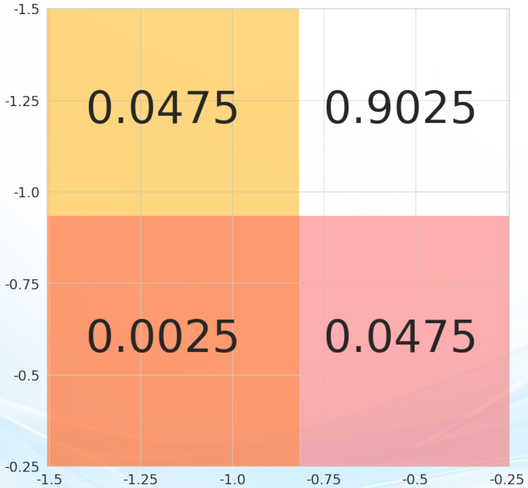
Частая ошибка: Выборки независимы $\Longrightarrow$ МПГ не нужна.



Вывод:

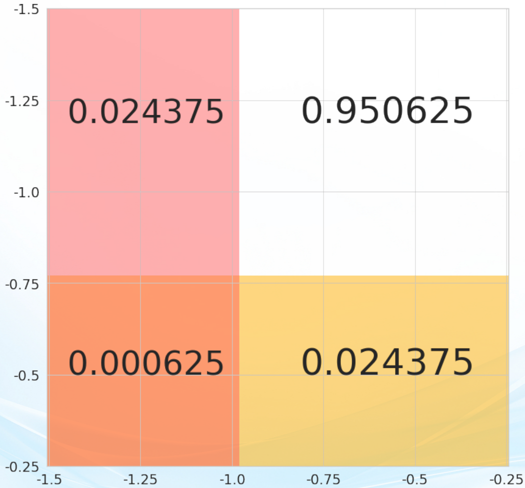
вероятность допустить хотя бы одну ошибку равна 0.0975, если обе основные гипотезы верны

Сравнение: без корректировки



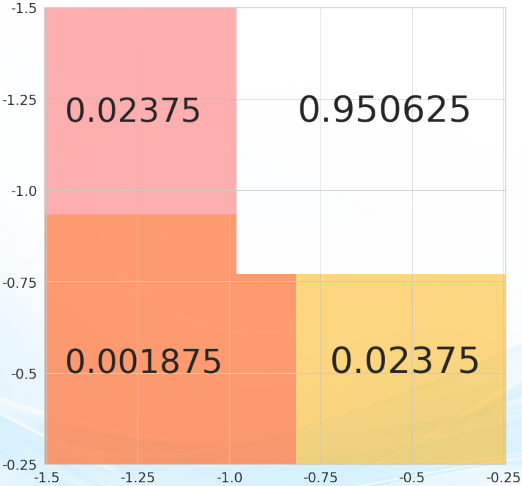
Вероятности указаны при справедливости H_1 и H_2

Сравнение: метод Бонферрони



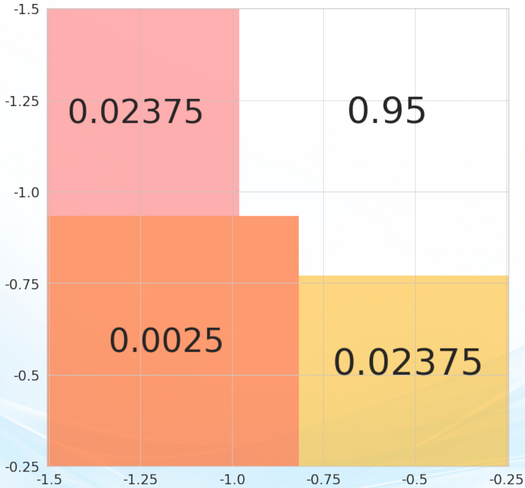
Вероятности указаны при справедливости H_1 и H_2

Сравнение: метод Холма



Вероятности указаны при справедливости H_1 и H_2

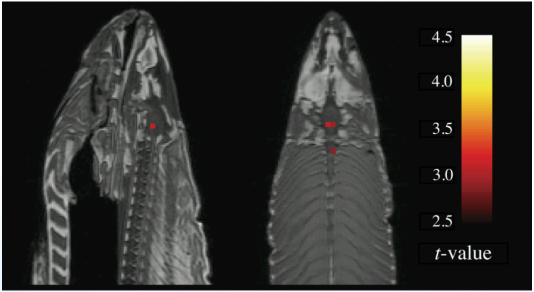
Сравнение: метод Бенджамини-Хохберга (не контр. FWER)



Вероятности указаны при справедливости H_1 и H_2

Удивительные открытия

2009 год. МРТ мозга мертвого самца лосося:



МРТ дает 3D-изображение на 130 000 вокселей.

Эксперимент: Лососю показывали фото и просили его пояснить, какие эмоции испытывают люди с картинки.

Обработка: Для каждого вокселя тестируется гипотеза о наличии активации этого участка мозга.

Удивительные открытия

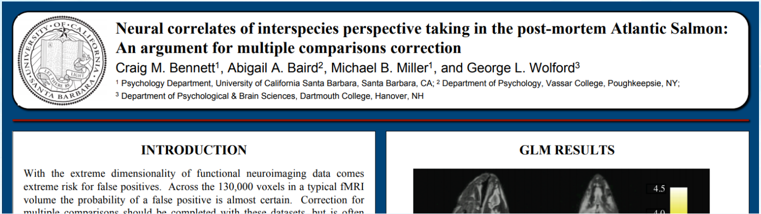
Результат: Для каждой картинки для нескольких вокселей мозга p-value оказывалось меньше 0.001.

Вывод: мертвый лосось реагирует на все!!!

Авторы удостоились Шнобелевской премии (2012 год)
за открытие в области неврологии.

При применении МПГ лосось переставал на что-либо реагировать...

http://prefrontal.org/files/posters/Bennett-Salmon-2009.pdf



Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction

Craig M. Bennett[1], Abigail A. Baird[2], Michael B. Miller[1], and George L. Wolford[3]

[1] Psychology Department, University of California Santa Barbara, Santa Barbara, CA; [2] Department of Psychology, Vassar College, Poughkeepsie, NY; [3] Department of Psychological & Brain Sciences, Dartmouth College, Hanover, NH

INTRODUCTION

With the extreme dimensionality of functional neuroimaging data comes extreme risk for false positives. Across the 130,000 voxels in a typical fMRI volume the probability of a false positive is almost certain. Correction for multiple comparisons should be completed with these datasets, but is often

GLM RESULTS


ВСЁ!